

Seeing the Forest from the Trees in Two Looks: Matrix Sketching by Cascaded Bilateral Sampling

Kai Zhang¹, Chuanren Liu², Jie Zhang³, Hui Xiong⁴, Eric Xing⁵, Jieping Ye⁶

¹NEC Laboratories Amercia, Princeton

²Lebow College of Business, Drexel University, Philadelphia

³Center of Computational Biology, Fudan University China

⁴Management Science & Information Systems Department, Rutgers

⁵Machine Learning Department, Carnegie Mellon University

⁶Department of Computational Medicine and Bioinformatics
University of Michigan, Ann Arbor

July 28, 2016

Abstract

Matrix sketching is aimed at finding close approximations of a matrix by factors of much smaller dimensions, which has important applications in optimization and machine learning. Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, state-of-the-art randomized algorithms take $\mathcal{O}(m \cdot n)$ time and space to obtain its low-rank decomposition. Although quite useful, the need to store or manipulate the entire matrix makes it a computational bottleneck for truly large and dense inputs. Can we sketch an m -by- n matrix in $\mathcal{O}(m + n)$ cost by accessing only a small fraction of its rows and columns, without knowing anything about the remaining data? In this paper, we propose the cascaded bilateral sampling (CABS) framework to solve this problem. We start from demonstrating how the approximation quality of bilateral matrix sketching depends on the encoding powers of sampling. In particular, the sampled rows and columns should correspond to the code-vectors in the ground truth decompositions. Motivated by this analysis, we propose to first generate a pilot-sketch using simple random sampling, and then pursue more advanced, “follow-up” sampling on the pilot-sketch factors seeking maximal encoding powers. In this cascading process, the rise of approximation quality is shown to be lower-bounded by the improvement of encoding powers in the follow-up sampling step, thus theoretically guarantees the algorithmic boosting property. Computationally, our framework only takes linear time and space, and at the same time its performance rivals the quality of state-of-the-art algorithms consuming a quadratic amount of resources. Empirical evaluations on benchmark data fully demonstrate the potential of our methods in large scale matrix sketching and related areas.

1 Introduction

Matrix sketching is aimed at finding close approximations of a matrix by using factors of much smaller dimensions, which plays important roles in optimization and machine learning [14, 11, 21, 17, 16, 18, 19, 12]. A promising tool to solve this problem is low-rank matrix decomposition, which approximates an input matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ by $\mathbf{A} \approx \mathbf{P}\mathbf{Q}^\top$, where \mathbf{P} and \mathbf{Q} have a low column rank $k \ll m, n$. Recent advances in randomized algorithms have made it state-of-the-art in low-rank matrix sketching or decomposition. For example, Frieze *et al.* [9] and Drineas *et al.* [5] proposed to use monte-carlo sampling to select informative rows or columns from a matrix; Mahoney and Drineas [19] proposed CUR matrix decomposition and used *statistical leverage scores* to perform sampling; Halko *et al.* [12] proposed to project the matrix to a lower-dimensional space and then compute the desired factorization [6, 3].

These algorithms compute approximate decompositions in $\mathcal{O}(m \cdot n)$ time and space, which is more efficient than a singular value decomposition using $\mathcal{O}(n^2 m)$ time and $\mathcal{O}(m \cdot n)$ space (if $m \geq n$). However, the whole input matrix must be fully involved in the computations, either in computing high-quality sampling probabilities [5, 19, 9, 26], or being compressed into a lower-dimensional space [12]. This can lead to potential computational and memory bottlenecks in particular for truly large and dense matrices.

Is it possible to sketch an m -by- n matrix in $\mathcal{O}(m+n)$ time and space, by using only a small number of its rows and columns and never knowing the remaining entries? To the best of our knowledge, this is still an open problem. Actually, the challenge may even seem unlikely to resolve at first sight, because by accessing such a small fraction of the data, resultant approximation can be quite inaccurate; besides, a linear amount of resources could hardly afford any advanced sampling scheme other than the uniform sampling; finally, approximation of a general, rectangular matrix with linear cost is much harder than that of a positive semi-definite (PSD) matrix such as kernel matrix [25, 8, 7].

To resolve this problem, in this paper we propose a cascaded bilateral sampling (CABS) framework. Our theoretical foundation is an innovative analysis revealing how the approximation quality of bilateral matrix sketching is associated with the encoding powers of sampling. In particular, selected columns and rows should correspond to representative code-vectors in the ground-truth embeddings. Motivated by this, we propose to first generate a pilot-sketch using simple random sampling, and then pursue more advanced exploration on this pilot-sketch/embedding seeking maximal encoding powers. In this process, the rise of approximation quality is shown to be lower-bounded by the improvement of encoding powers through the follow-up sampling, thus theoretically guarantees the algorithmic boosting property. Computationally, both rounds of sampling-and-sketching operations require only a linear cost; however, when cascaded properly, the performance rivals the quality of state-of-the-art algorithms consuming a quadratic amount of resources. The CABS framework is highly memory and pass efficient by only accessing twice a small number of specified rows and columns, thus quite suitable to large and dense matrices which won't fit in the main memory. In the meantime, the sketching results are quite easy to interpret.

2 Related Work

2.1 Quadratic-Cost Algorithms

We first review state-of-the-art randomized algorithms. They typically consume quadratic, $\mathcal{O}(mn)$ time and space due to the need to access and manipulate the entire input matrix in their calculations.

Monte-Carlo sampling method [9, 5] computes approximate singular vectors of a matrix \mathbf{A} by selecting a subset of its columns using non-uniform, near-optimal sampling probabilities. The selected columns are re-scaled by the probabilities, and its rank- k basis \mathbf{Q}_k is then used to obtain the final decomposition. If the probabilities are chosen as $p_j \geq \beta \cdot \|\mathbf{A}_{[:,j]}\|^2 / (\sum_{i=1}^n \|\mathbf{A}_{[:,i]}\|^2)$, then with probability at least $1 - \delta$, one has $\|\mathbf{A} - \mathbf{Q}_k \mathbf{Q}_k^\top \mathbf{A}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \epsilon \|\mathbf{A}\|_F^2$, where \mathbf{A}_k is the best rank- k approximation, $\epsilon = 2\eta / \sqrt{k/\beta c}$, with $\beta \leq 1$, $\eta = 1 + \sqrt{(8/\beta) \log(1/\delta)}$.

Random projection methods [12] project a matrix \mathbf{A} into a subspace $\mathbf{Q} \in \mathbb{R}^{m \times k}$ with orthonormal columns such that $\mathbf{A} \approx \mathbf{Q} \mathbf{Q}^\top \mathbf{A}$. Computing \mathbf{Q} requires multiplying \mathbf{A} with a Gaussian test matrix (or random Fourier transform) $\mathbf{\Omega}$ with q steps of power iterations, $\mathbf{Y} = (\mathbf{A} \mathbf{A}^\top)^q \mathbf{A} \mathbf{\Omega}$, and then computing QR-decomposition $\mathbf{Y} = \mathbf{Q} \mathbf{R}$. Using a template $\mathbf{\Omega} \in \mathbb{R}^{m \times (k+p)}$ with over-sampling parameter p , $\mathbb{E} \|\mathbf{A} - \mathbf{Q} \mathbf{Q}^\top \mathbf{A}\| = [1 + 4\sqrt{(k+p) \cdot \min(m, n)/(p-1)}] \Sigma_{k+1}$, where \mathbb{E} is expectation with $\mathbf{\Omega}$, Σ_{k+1} is the $(k+1)$ th singular value. New projection methods are explored in [6, 3].

CUR matrix decomposition [19] was invented by Mahoney and Drineas to improve the interpretability of low-rank matrix decomposition. As shown in Figure 1, it samples a subset of columns and rows from \mathbf{A} , as $\mathbf{C} \in \mathbb{R}^{m \times k_c}$ and $\mathbf{R} \in \mathbb{R}^{k_r \times n}$, and then solve

$$\min_{\mathbf{U}} \|\mathbf{A} - \mathbf{C} \mathbf{U} \mathbf{R}\|_F^2 \rightarrow \mathbf{U}^* = \mathbf{C}^\dagger \mathbf{A} \mathbf{R}^\dagger, \quad (1)$$

where † is pseudo-inverse. The CUR method preserves the sparsity and non-negativity properties of input matrices. If leverage scores are used for sampling, then with high probability, $\|\mathbf{A} - \mathbf{C} \mathbf{U} \mathbf{R}\|_F \leq (2 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F$. The leverage scores are computed by top- r left/right singular vectors, taking $\mathcal{O}(mn)$ space and $\mathcal{O}(mnr)$ time. Computing \mathbf{U} in (1) involves multiplying \mathbf{A} with \mathbf{C}^\dagger and \mathbf{R}^\dagger . Therefore, even using simple random sampling, CUR may still take $\mathcal{O}(mn)$ time and space.

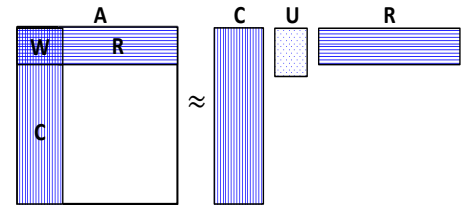


Figure 1: Illustration of CUR method.

2.2 Linear-Cost Algorithms

The randomized algorithms discussed in Section 2.1 take at least quadratic time and space. One way to reduce the cost to a linear scale is to restrict the calculations to only a small fraction of rows and columns. In the literature, related

work is quite limited. We use a general variant of the CUR method, called *Bilateral Re-sampling CUR* (BR-CUR), to initiate the discussion.

Algorithm 1: Bilateral Resampling CUR (BR-CUR)

Input: \mathbf{A} , base sampling $\mathcal{I}_r^b, \mathcal{I}_c^b$, target sampling $\mathcal{I}_r^t, \mathcal{I}_c^t$;

Output: $\mathbf{C}, \mathbf{U}, \mathbf{R}$

- 1: Compute bases $\mathbf{C} = \mathbf{A}_{[:, \mathcal{I}_c^b]}$, $\mathbf{R} = \mathbf{A}_{[\mathcal{I}_r^b, :]}$.
 - 2: Sample on bases $\bar{\mathbf{C}} = \mathbf{C}_{[\mathcal{I}_r^t, :]}$, $\bar{\mathbf{R}} = \mathbf{R}_{[:, \mathcal{I}_c^t]}$.
 - 3: Compute target block $\mathbf{M} = \mathbf{A}_{[\mathcal{I}_r^t, \mathcal{I}_c^t]}$.
 - 4: Solve $\mathbf{U}^* = \arg \min_{\mathbf{U}} \|\mathbf{M} - \bar{\mathbf{C}}\mathbf{U}\bar{\mathbf{R}}\|_F^2$.
 - 5: Reconstruct by $\mathbf{A} \approx \mathbf{C}\mathbf{U}^*\mathbf{R}$.
-

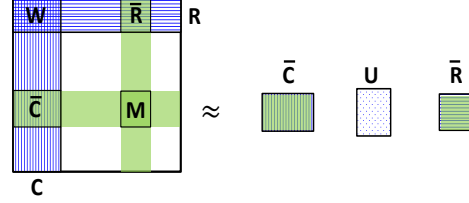


Figure 2: Illustration of BR-CUR.

As illustrated in Figure 2, BR-CUR has two rounds of samplings: blue “base”-sampling \mathcal{I}_r^b (row) and \mathcal{I}_c^b (column) to construct \mathbf{C} and \mathbf{R} , and green “target”-sampling \mathcal{I}_r^t (row) and \mathcal{I}_c^t (column) to specify a sub-matrix from \mathbf{A} . In computing \mathbf{U} (step 4), it only attempts to minimize the approximation error on the target sub-matrix, therefore being computationally quite efficient. Assume k_1 and k_2 are the number of random samples selected in the base and target sampling, respectively. Then BR-CUR only takes $\mathcal{O}((m+n)(k_1+k_2))$ space and $\mathcal{O}((m+n)\max(k_1, k_2)^2)$ time.

The BR-CUR procedure subsumes most linear-cost algorithms for matrix sketching in the literature.

1. Nyström method [25, 8, 7]: in case of symmetric positive semi-definite matrix \mathbf{A} , upon using the same base and target sampling, and the same row and column sampling, $\mathcal{I}_c^b = \mathcal{I}_r^b = \mathcal{I}_c^t = \mathcal{I}_r^t$.
2. Pseudo-skeleton [10] or bilateral projection [29]: in case of rectangular matrix \mathbf{A} , and using the same base and target sampling, i.e., $\mathcal{I}_r^b = \mathcal{I}_r^t$ and $\mathcal{I}_c^b = \mathcal{I}_c^t$. It generalizes the Nyström method from symmetric to rectangular matrices. Let \mathbf{W} be the intersection of \mathbf{C} and \mathbf{R} , then it has compact form,

$$\mathbf{A} \approx \mathbf{C}\mathbf{W}^\dagger\mathbf{R}. \quad (2)$$

3. Sketch-CUR [24]: in case of rectangular \mathbf{A} and independent base/target sampling (different rates).

These algorithms only need to access a small fraction of rows and columns to reconstruct the input matrix. However, the performance can be inferior, in particularly on general, rectangular matrices. Full discussions are in Section 4.2. More recently, online streaming algorithms are also designed for matrix sketching [27, 15, 20]. Their memory cost is much smaller, but the whole matrix still needs to be fully accessed and the time complexity is at least $\mathcal{O}(m \cdot n)$. For sparse matrices, the rarity of non-zeros entries can be exploited to design algorithms in input sparsity time [4]. Note that the method proposed in this paper is applicable to both dense and sparse matrices; in particular, significant performance gains will be observed in both scenarios in our empirical evaluations (Section 5).

3 Theoretic Analysis of Bilateral Matrix Sketching

In this section, we present a novel theoretic analysis on bilateral sketching of general, rectangular matrices. It links the quality of matrix sketching with the encoding powers of bilateral sampling, and inspires a weighted k -means procedure as a novel bilateral sampling scheme.

3.1 A New Error Bound

In the literature, most error bound analysis for matrix low-rank approximation is of the following theme: a sampling scheme is pre-determined and then probabilistic theoretical guarantees are derived, typically on how many samples should be selected to achieve a desired accuracy [12, 5, 19, 9, 23]. In this work, our goal is quite different: *we want to maximally reduce the approximation error given a fixed rate of sampling*. Therefore, our error bound is expressed in terms of the numerical properties of sampling, instead of a specific sampling scheme. Such a heuristic error bound will shed more light on the design of sampling schemes to fully exploit a limited computing resource, thus particularly useful to practitioners.

Given an input matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, assume $\mathbf{A} = \mathbf{P}\mathbf{Q}^\top$, where $\mathbf{P} \in \mathbb{R}^{m \times r}$ and $\mathbf{Q} \in \mathbb{R}^{n \times r}$ are exact decomposition. Without loss of generality suppose we select k columns $\mathbf{C} = \mathbf{A}_{[:, \mathcal{Z}^c]}$ and k rows $\mathbf{R} = \mathbf{A}_{[\mathcal{Z}^r, :]}$, where \mathcal{Z}^c and \mathcal{Z}^r are sampling indices. These indices locate representative instances (rows) in \mathbf{P} and \mathbf{Q} , denoted by $\mathbf{Z}^r = \mathbf{P}_{[\mathcal{Z}^r, :]}$ and $\mathbf{Z}^c = \mathbf{Q}_{[:, \mathcal{Z}^c]}$, respectively. In order to study how the bilateral sampling affects matrix sketching result, we adopt a clustered data model. Let the m rows of \mathbf{P} be grouped to k clusters, where the cluster representatives are rows in \mathbf{Z}^r ;

similarly, let the n rows in \mathbf{Q} be grouped into k clusters, where the cluster representatives are rows in \mathbf{Z}^c . Let the $s^r(i)$ be the cluster assignment function that maps the i -th row in \mathbf{P} to the $s^r(i)$ -th row in \mathbf{Z}^r ; similarly, $s^c(i)$ maps the i -th row in \mathbf{Q} to the $s^c(i)$ -th row in \mathbf{Z}^c . Then, the errors of reconstructing \mathbf{P} and \mathbf{Q} using the representatives in \mathbf{Z}^r and \mathbf{Z}^c via respective mapping function $s^r(\cdot)$ and $s^c(\cdot)$ can be defined as

$$e^r = \sum_{l=1}^m \|\mathbf{P}_{[l,:]} - \mathbf{Z}_{[s^r(l),:]}^r\|^2, \quad e^c = \sum_{l=1}^n \|\mathbf{Q}_{[l,:]} - \mathbf{Z}_{[s^c(l),:]}^c\|^2. \quad (3)$$

We also define T^r and T^c as the maximum cluster sizes in \mathbf{P} and \mathbf{Q} , respectively, as

$$T^r = \max_{1 \leq y \leq k} |\{i : s^r(i) = y\}|, \quad T^c = \max_{1 \leq y \leq k} |\{i : s^c(i) = y\}|. \quad (4)$$

Given these preliminaries, we can then analyze how the matrix sketching error is associated with the encoding powers of bilateral sampling (in reconstructing the decompositions \mathbf{P} and \mathbf{Q}) as follows.

Theorem 1 *Given an input matrix \mathbf{A} , and suppose one samples k columns \mathbf{C} and k rows \mathbf{R} , with the intersection \mathbf{W} . Then we can bound the approximation error (2) as follows.*

$$\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{R}\|_F \leq \left(\sqrt{6k\theta T^{\frac{3}{2}}}\right) \cdot \sqrt{e^r + e^c} + (k\theta T \|\mathbf{W}^\dagger\|_F) \cdot \sqrt{e^c e^r} \quad (5)$$

Here $T = \max(T^r, T^c)$, T^r and T^c are defined in (4), e^r and e^c are defined in (3); and θ is a data dependent constant. Proof of Theorem 1 can be found in Section 1 of supplementary material.

From Theorem 1, we can see that given a fixed sampling rate k (so T and $\|\mathbf{W}^\dagger\|_F$ will more or less remain the same too), the key quantities affecting the matrix sketching error are e^r and e^c (3), the encoding errors of reconstructing \mathbf{P} and \mathbf{Q} with their representative rows whose indices are specified in the bilateral sampling. In case both e^c and e^r approach zero, the sketching error will also approach zero. Namely, choosing a bilateral sampling that can reduce the encoding errors (3) is an effective way to bound matrix sketching error. We visualized their relations in Figure 4. Here, given \mathbf{A} with decomposition $\mathbf{P}\mathbf{Q}^\top$, we perform bilateral samplings many times, each time using an arbitrary choice such as uniform sampling, vector quantization (with different number of iterations), and so on, such that the resultant encoding errors vary a lot. Then we plot the encoding errors (e^r, e^c) and color-code it with the corresponding matrix sketching error \mathcal{E} . As can be seen, \mathcal{E} shows a clear correlation with e^c and e^r . Only when both e^c and e^r are small (blue), \mathcal{E} will be small; or else if either e^c or e^r is large, \mathcal{E} will be large too.

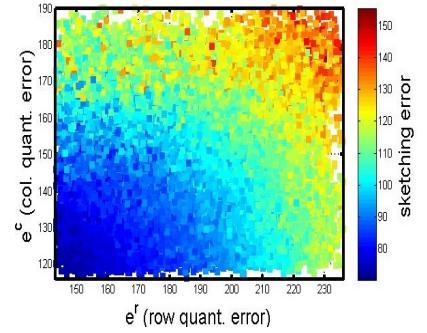


Figure 4: Matrix sketching error vs. row and column encoding errors.

3.2 Sampling on Low-rank Embeddings: Weighted k -means

Theorem 1 provides an important criterion for sampling: the selected rows and columns of \mathbf{A} should correspond to representative code-vectors in the low-rank embeddings \mathbf{P} and \mathbf{Q} , in order for the sketching error to be well bounded. To achieve this goal, we propose to use k -means sampling independently on \mathbf{P} and \mathbf{Q} , which can quickly reduce their respective encoding errors in just a few iterations. Of course, exact decompositions \mathbf{P} and \mathbf{Q} are impractical and possibly high dimensional. Therefore, we resort to an alternative low-dimensional embedding $\mathbf{P} \in \mathbb{R}^{m \times k}$, $\mathbf{Q} \in \mathbb{R}^{n \times k}$ that will be discussed in detail in the cascaded sampling framework in Section 4.

Here we first study the performance of k -means sampling. We note that dense and sparse matrices have different embedding profiles. Energy of dense matrices spreads across rows and columns, so the embedding has a fairly uniform distribution (Figure 5(a)). For sparse matrices whose entries are mostly zeros, the embedding collapses towards the origin (Figure 5(b)). The k -means algorithm assigns more centers in densely distributed regions. Therefore the clustering centers are uniform for dense matrices (Figure 5(a)), but will be attracted to the origin for sparse matrices (Figure 5(b)). These observations inspire us to perform an importance-weighted k -means sampling as follows

$$e^r = \sum_{l=1}^m \|\mathbf{P}_{[l,:]} - \mathbf{Z}_{[s^r(l),:]}^r\|^2 \cdot \Upsilon(\|\mathbf{P}_{[l,:]} \|_2), \quad e^c = \sum_{l=1}^n \|\mathbf{Q}_{[l,:]} - \mathbf{Z}_{[s^c(l),:]}^c\|^2 \cdot \Upsilon(\|\mathbf{Q}_{[l,:]} \|_2). \quad (6)$$

Here we use the norm of $\mathbf{P}_{[l,:]}$ (or $\mathbf{Q}_{[l,:]}$) to re-weight the objective of k -means in (3), because it is an upper-bound of the energy of the i th row in \mathbf{A} (up to a constant scaling), as $\|\mathbf{A}_{[i,:]} \| = \|\mathbf{P}_{[i,:]} \cdot \mathbf{Q}\| \leq \|\mathbf{P}_{[i,:]} \| \cdot \|\mathbf{Q}\|$. The $\Upsilon(\cdot)$

is a monotonous function adjusting the weights (e.g., power, sigmoid, or step function). Here, priority is given to rows/columns with higher energy, and as a result the k -means cluster centers will then be pushed away from the origin (Figure 5(c)). In practice, we will chose a fast-growing function Υ for sparse matrices, and a slowly-growing (or constant) function Υ for dense matrices, and any k -means clustering center will be replaced with its closest in-sample point. Finally, the weighing can be deemed a prior knowledge (preference) on approximating rows and columns of the input matrix, which does not affect the validity of Theorem 1.

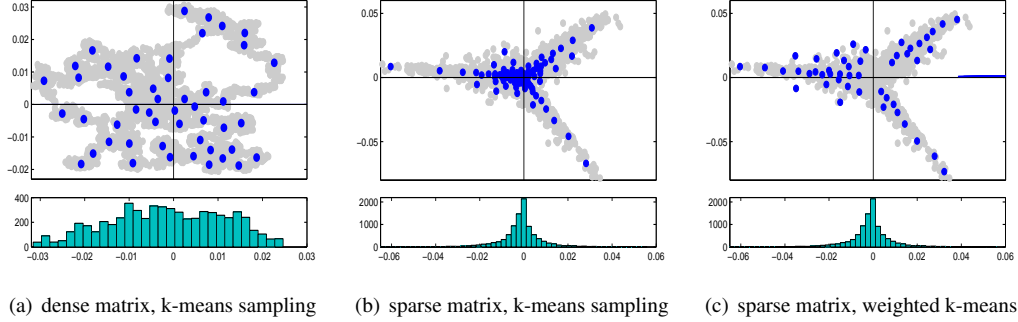


Figure 5: Top-2 dimensions of embedding \mathbf{P} , histogram on horizontal dimension, and (weighted) k -means.

4 Cascaded Bilateral Sketching (CABS) Framework

Theorem 1 suggests that one perform weighted k -means sampling (6) on the bilateral embeddings of the input matrix to effectively control the approximation error. Since computing an exact embedding is impractical, we will resort to approximate embeddings discussed in the following framework.

Algorithm 2: Cascaded Bilateral Sampling (CABS)

Input: \mathbf{A} ; **Output:** $\mathbf{A} \approx \bar{\mathbf{U}}\bar{\mathbf{S}}\bar{\mathbf{V}}^\top$

- 1: *Pilot Sampling*: randomly select k columns and k rows $\mathbf{C} = \mathbf{A}_{[:, \mathcal{I}^c]}$, $\mathbf{R} = \mathbf{A}_{[\mathcal{I}^r, :]}$, $\mathbf{W} = \mathbf{A}_{[\mathcal{I}^r, \mathcal{I}^c]}$.
 - 2: *Pilot Sketching*: run $[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \text{sketching}(\mathbf{C}, \mathbf{R}, \mathbf{W})$, let $\mathbf{P} = \mathbf{U}\mathbf{S}^{\frac{1}{2}}$, and $\mathbf{Q} = \mathbf{V}\mathbf{S}^{\frac{1}{2}}$.
 - 3: *Follow-up sampling*: perform weighted k -means on \mathbf{P} and \mathbf{Q} , respectively, to obtain row index $\bar{\mathcal{I}}^r$ and column index $\bar{\mathcal{I}}^c$; let $\bar{\mathbf{C}} = \mathbf{A}_{[:, \bar{\mathcal{I}}^c]}$, $\bar{\mathbf{R}} = \mathbf{A}_{[\bar{\mathcal{I}}^r, :]}$, and $\bar{\mathbf{W}} = \mathbf{A}_{[\bar{\mathcal{I}}^r, \bar{\mathcal{I}}^c]}$.
 - 4: *Follow-up sketching*: run $[\bar{\mathbf{U}}, \bar{\mathbf{S}}, \bar{\mathbf{V}}] = \text{sketching}(\bar{\mathbf{C}}, \bar{\mathbf{R}}, \bar{\mathbf{W}})$.
-

The CABS framework has two rounds of operations, each round with a sampling and sketching step. In the first round, we perform a simple, random sampling (step 1) and then compute a pilot sketching of the input matrix (step 2). Although this pilot sketching can be less accurate, it provides a compact embedding of the input matrix (\mathbf{P} and \mathbf{Q}). In the follow-up round, as guided by Theorem 1, we then apply weighted k -means on \mathbf{P} and \mathbf{Q} to identify representative samples (step 3); resultant sampling is used to compute the final sketching result (step 4). As will be demonstrated both theoretically (Section 4.2) and empirically (Section 5), it is exactly this follow-up sampling that allows us to extract a set of more useful rows and columns, thus significantly boosting the sketching quality.

The `sketching` routine computes the decomposition of a matrix using only selected rows and columns, as we shall discuss in Section 4.1. As a result, CABS takes only $\mathcal{O}((m+n)(k_1+k_2))$ space and $\mathcal{O}((m+n)k_1k_2c)$ time, where k_1 and k_2 is the pilot and follow-up sampling rate, respectively, and c is the number of k -means iterations. In practice, $k_1 = k_2 \ll m, n$ and $c = 5$ in all our experiment, so the complexities are linear in $m+n$. The decomposition is also quite easy to interpret because it is expressed explicitly with a small subset of representative rows and columns.

4.1 The sketching routine

In this section we discuss the `sketching` routine used in Algorithm 2. Given a subset of rows \mathbf{R} , columns \mathbf{C} , and their intersection \mathbf{W} from a matrix \mathbf{A} , this routine returns the decomposition $\mathbf{A} \approx \mathbf{U}\mathbf{S}\mathbf{V}^\top$. Both Sketch-CUR and Pseudo-skeleton method are possible candidates, however, in practice they can be numerically sensitive. As shown in

Figure 6, their approximation error varies with the number of singular vectors used in computing the pseudo-inverse. The optimal number can be small and different from data to data (w.r.t. matrix size). Another observation is that, Pseudo-skeleton is always superior to Sketch-CUR when they use the same sampling rates.

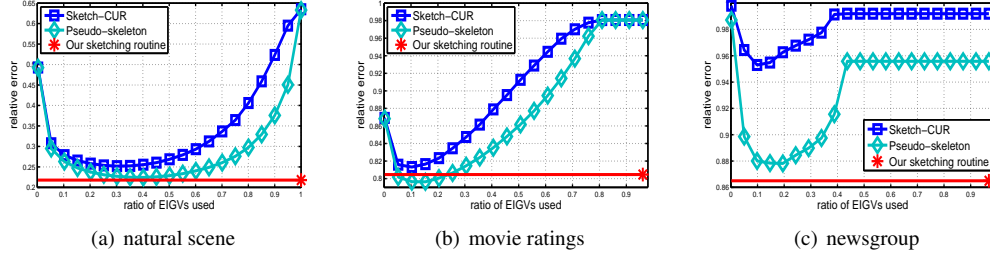


Figure 6: Performance of linear-cost algorithms is sensitive to the number of singular vectors used.

In the following we propose a more stabilized variant of Pseudo-skeleton method (Equation 2). Assuming an SVD $\mathbf{W} = \mathbf{U}_w \Sigma_w \mathbf{V}_w^\top$, then $\mathbf{W}^\dagger = \mathbf{V}_w \Sigma_w^{-1} \Sigma_w \Sigma_w^{-1} \mathbf{U}_w^\top$. Plug this into $\mathbf{A} \approx \mathbf{C} \mathbf{W}^\dagger \mathbf{R}$, we have

$$\mathbf{A} \approx (\mathbf{C} \mathbf{V}_w \Sigma_w^{-1}) \Sigma_w (\Sigma_w^{-1} \mathbf{U}_w^\top \mathbf{R}).$$

Here, \mathbf{U}_w and \mathbf{V}_w are left and right singular vectors of \mathbf{W} , extrapolated via $\mathbf{U}_w^\top \mathbf{R}$ and $\mathbf{C} \mathbf{V}_w$, respectively, and then normalized by the singular values Σ_w . In case $\Sigma_w(i, i)$ approaches zero, the normalization becomes numerically unstable. To avoid this ambiguity, we propose to use the norms of the extrapolated singular-vectors for normalization, as

$$\mathbf{A} \approx (\mathbf{C} \mathbf{V}_w \mathbf{N}_c^{-1}) \sqrt{\frac{mn}{k^2}} \Sigma_w (\mathbf{N}_r^{-1} \mathbf{U}_w^\top \mathbf{R}), \quad \text{s.t. } \mathbf{N}_c = \text{diag}(\|\mathbf{C} \mathbf{V}_w\|_\otimes), \mathbf{N}_r = \text{diag}(\|\mathbf{R}^\top \mathbf{U}_w\|_\otimes)$$

Here $\text{diag}(\cdot)$ fills a diagonal matrix with given vector, $\|\cdot\|_\otimes$ returns column-wise norms, namely \mathbf{N}_r and \mathbf{N}_c are norms of extrapolated singular vectors. The constant \sqrt{mn}/k adjusts the scale of solution. We can then define *sketching* routine with $\mathbf{U} = \mathbf{C} \mathbf{V}_w \mathbf{N}_c^{-1}$, $\mathbf{V} = \mathbf{N}_r^{-1} \mathbf{U}_w^\top \mathbf{R}$, and $\Sigma = \Sigma_w \sqrt{mn}/k$. As can be seen from Figure 6, it gives stable result by using all singular vectors. The solution can be orthogonalized in linear cost [12] (see Section 3 in supplementary material). In practice, one can also use a validation set to choose the optimal number of singular vectors.

One might wonder how previous work tackled the instability issue, however, related discussion is hard to find. In the literature, expensive computations such as power iteration [29] or high sampling rate [24] were used to improve the performance of Pseudo-skeleton and Sketch-CUR, respectively. Neither suits our need. In comparison, our solution requires no extra computations or sampling.

In positive semi-definite (PSD) matrices, intriguingly, numerical sensitivity diminishes by sampling the same subset of rows and columns, which reduces to the well-known Nyström method [25, 8, 7] (see Section 2 of supplementary material for detail). Namely, Nyström method is numerically stable on PSD matrices, but its generalized version is more sensitive on rectangular matrices. We speculate that PSD matrix resides in a Riemannian manifold [1], so symmetric sampling better captures its structure, however rectangular matrix is not endowed with any structural constraint. This makes the approximation of rectangular matrices particularly challenging compared with that of PSD matrices, and abundant results of the Nyström method may not be borrowed here directly [25, 8, 13, 7, 28]. By stabilizing the *sketching* routine and employing it in the cascaded sampling framework, CABS will be endowed with superior performance in sketching large rectangular matrices.

4.2 Algorithmic Boosting Property of CABS

In the literature, many two-step methods were designed for fast CUR decomposition. They start from an initial decomposition such as fast JL-transform [6] or random projection [2, 22, 24], and then compute a sampling probability with it for subsequent CUR [2, 22, 23]. We want to emphasize the differences between CABS and existing two-step methods. *Algorithmically*, CABS only accesses a small part of the input matrix; two-step methods often need to manipulate the whole matrix. Moreover, CABS performs the follow-up sampling on the bilateral low-rank embeddings,

which are compact, multivariate summary of the input matrix; in two-step methods, the sampling probability is obtained by reducing the initial decomposition to a univariate probability scores. *Conceptually*, CABS is targeted on algorithmically boosting the performance of cheap sampling routines. In comparison, two-step methods only focus on theoretic performance guarantees of the second (final) step, and less attention was put on the performance gains from the first step to the second step.

How to quantify the rise (or drop) of approximation quality in a two-step method? How to choose the right pilot and follow-up sampling to save computational costs and maximize performance gains? These are fundamental questions we aim to answer. For the second question, we address it by creatively cascading random sampling with a weighted k -means, thus deriving a working example of “algorithmic boosting”. One might want to try pairing existing sketching/sampling algorithms to achieve the same goal, however, choosing an effective follow-up sampling from existing strategies can be non-trivial (see Section 5 for detailed discussion). For the first question, a thorough theoretic answer is quite challenging, nevertheless, we still provide a useful initial result. We show that, the decrement of the error bound from the pilot sketching to the follow-up sketching in CABS, is lower bounded by the drop of encoding errors achieved through the follow-up sampling of k -means. In other words, the better the encoding in the follow-up sampling step, the larger the performance gain.

Theorem 2 *Let the error bound of the pilot and follow-up sketching in CABS be Ψ_p and Ψ_f , respectively. Then the error bound will drop by at least the following amount*

$$\Psi_p - \Psi_f \geq \sqrt{\frac{3k\theta T_p^c T_p^r (T_p^r + T_p^c)}{2(e_p^r + e_p^c)}} |\Delta_e^r + \Delta_e^c| + k\theta \|\mathbf{W}_p^\dagger\|_F \sqrt{\frac{T_p^c T_p^r}{e_p^r e_p^c}} |\Delta_e^r e_f^c + \Delta_e^c e_f^r + \Delta_e^r \Delta_e^c|.$$

Parameters are defined in the same way as in Theorem 1, and sub-index $\{p, f\}$ denotes pilot or follow-up; Δ_e^r (Δ_e^c) is the drop of row (column) encoding error in the follow-up k -means sampling. Proof of Theorem 2 can be found in Section 4 of supplementary material.

The algorithmic boosting effect is not dependent on the sketching routine adopted in Algorithm 2. Empirically, by using less stable routines (Pseudo-skeleton or Sketch-CUR), significant performance gains are still observed. Namely CABS is a general framework to achieve algorithmic boosting. In practice, the superior performance of CABS is attributed to both reliable sketching routine and the cascading/boosting mechanism.

5 Experiments

All experiments run on a server with 2.6GHZ processor and 64G RAM. Benchmark data sets are described in Table 1. All codes are written in matlab and fully optimized by vectorized operations and matrix computations, to guarantee a fair comparison in case time consumption is considered.

Table 1: Summary of benchmark data sets.

| data | #row | #column | sparsity | source urls |
|---------------|-------|---------|----------|---|
| movie ratings | 27000 | 46000 | 0.56% | http://grouplens.org/datasets/movielens/ |
| newsgroup | 18774 | 61188 | 0.22% | http://kdd.ics.uci.edu/databases/20newsgroups/ |
| natural scene | 6480 | 7680 | dense | http://wallpapershome.com/ |
| hubble image | 15852 | 12392 | dense | http://hubblesite.org/gallery/ |

First we compare all linear-cost algorithms, including: (a) Sketch-CUR [24], where the target sampling rate is chosen three times as much as the base sampling rate; (b) Pseudo-skeleton [10], which is the generalized Nyström method; (c) Pilot-sketch, which is step 2 of Algorithm 2; (d) Followup-sketch (w-kmeans), which is step 4 of Algorithm 2; (e) Followup-sketch (leverage), a variant of our approach using approximate leverage scores for the follow-up sampling; (f) Followup-sketch (hard-thrhd), a variant of our approach using top- k samples with largest weighting coefficients (Equation 6) for the follow-up sampling. In Figure 7(a), we gradually increase the number of selected rows and columns from 1% to 10%, and then report the averaged error $\|\mathbf{A} - \hat{\mathbf{A}}\|_F / \|\mathbf{A}\|_F$ over 20 repeats. For simplicity, the number of selected rows and columns are both k , and the sampling rate is defined as k/\sqrt{mn} . Our observations

are as follows: (1) our pilot sketching is already more accurate than both Pseudo-skeleton and Sketch-CUR; (2) our follow-up sketching result using the weighted k -means sampling strategy is consistently and significantly more accurate than our pilot sketching, which clearly demonstrates the “algorithmic boosting” effect of our framework; (3) on using leverage scores for the follow-up sampling, the performance gain is lower than the weighted k -means strategy, and becomes insignificant on dense matrices; we speculate that on dense matrices, since leverage scores are more uniformly distributed (Figure 5(a)), they are less discriminative and can introduce redundancy when used as sampling probabilities; (4) using hard-threshold sampling on the weighting coefficients is particularly beneficial on sparse matrices, but the results degenerate on dense matrices and so are skipped. This is because the weighting coefficients are clear indicators of the importance (or energy) of the rows or columns, and can be very discriminative on sparse matrices (see Figure 5(b)); on dense matrices, all the rows and columns can be equally important, therefore the weighting coefficients are no longer informative.

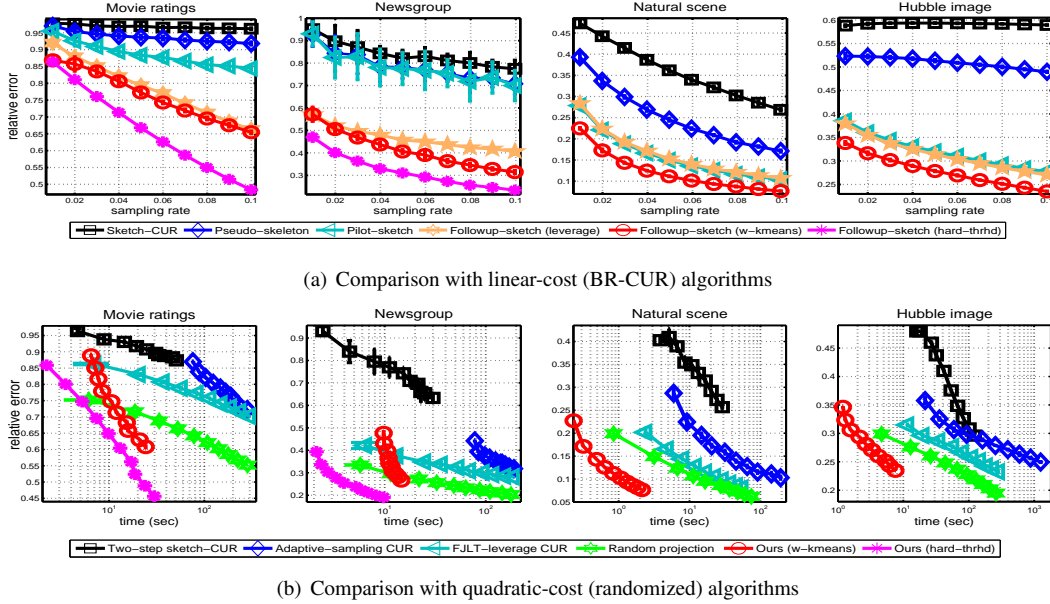


Figure 7: Sketching performance of our method and competing methods.

Then we compare our approach with state-of-the-art randomized algorithms (all with quadratic costs), including (a) Two-step Sketch-CUR, which first performs random projection [12] and then Sketch-CUR [22]; (b) Adaptive sampling CUR [23], which uses the error distribution of an initial approximation to guide extra sampling; (c) FJLT-leverage CUR, which uses fast JL-transform for initial decomposition and then perform CUR [6]; (d) Random projection [12], with $q = 1$ step of power iteration; (e) Ours (w-kmeans); and (f) Ours (hard-thrhd) for sparse matrices. Each algorithm is repeated 20 times and averaged approximation error over time consumption is reported in Figure 7(b). We have the following observations: (1) Random projection is very accurate but the time consumption can be significant; (2) FJLT-leverage CUR has similar time consumption and can be less accurate; (3) Adaptive-sampling CUR is computationally most expensive (in computing the residue of approximation); (4) Two-step Sketch-CUR is the least accurate, which we speculate is due to the instability of Sketch-CUR; (5) Our approach with weighted k -means sampling has a clear computational gain in dense matrices with a good accuracy; (6) Our approach using hard-threshold sampling performs particularly well in sparse matrices. Overall, our approaches have competing accuracies but significantly lower time and space consumption. The larger the input matrices, the higher the performance gains that can be expected.

6 Conclusion and Future Work

In this paper, we propose a novel computational framework to boost the performance of cheap sampling and sketching routines by creatively cascading them together. Our methods is particularly time and memory efficient, and delivers promising accuracy that matches with state-of-the-art randomized algorithms. In the future, we will pursue more

general theoretic guarantees and delineations; we are also applying the framework in parallel environment with some promising preliminary results.

References

- [1] R. Bhatia. *Positive Definite Matrices*. Princeton University Press, 2015.
- [2] C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near optimal column-based matrix reconstruction. In *IEEE Annual Symposium on Foundations of Computer Science*, pages 305–314, 2011.
- [3] K.L. Clarkson, P. Drineas, M. Magdon-Ismail, M.W. Mahoney, X. Meng, and D.P. Woodruff. The fast cauchy transform and faster robust linear regression. In *Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 466–477, 2013.
- [4] K.L. Clarkson and D.P. Woodruff. Low rank approximation and regression in input sparsity time. In *Annual ACM symposium on Theory of computing*, pages 81–90, 2013.
- [5] P. Drineas, R. Kannan, and M.W. Mahoney. Fast monte carlo algorithms for matrices II: computing a low-rank approximation to a matrix. *SIAM Journal of Computing*, 36(1):158–183, 2006.
- [6] P. Drineas, M. Magdon-Ismail, M.W. Mahoney, and D.P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(1):3475–3506, 2012.
- [7] P. Drineas and M.W. Mahoney. On the Nystrom method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.
- [8] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystrom method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.
- [9] A. Frieze, R. Kannan, and S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM*, 51(6):1025–1041, 2004.
- [10] S.A. Goreinov, E.E. Tyrtyshnikov, and N.L. Zamarashki. A theory of pseudoskeleton approximations. *Linear Algebra and its Applications*, 261(1):1–21, 1997.
- [11] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.
- [12] N. Halko, P.G. Martinsson, and J.A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.
- [13] S. Kumar, M. Mohri, and A. Talwalkar. Sampling methods for the nystrom method. *Journal of Machine Learning Research*, 13(1):981–1006, 2012.
- [14] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788 – 791, 1999.
- [15] E. Liberty. Simple and deterministic matrix sketching. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 581–588, 2013.
- [16] E. Liberty, F. Woolfe, P. Martinsson, V. Rokhlin, and M. Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of National Academy of Sciences*, 104(51):20167–20172, 2007.
- [17] U.V. Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586, 2007.
- [18] M.W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123 – 224, 2011.
- [19] M.W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of National Academy of Sciences*, 697–702(106):3, 2009.
- [20] I. Mitliagkas, C. Caramanis, and P. Jain. Memory limited, streaming pca. In *Advances in Neural Information Processing Systems*, pages 2886–2894, 2013.
- [21] Masashi Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *International conference on Machine learning*, pages 905–912, 2006.
- [22] S. Wang and Z. Zhang. A scalable cur matrix decomposition algorithm: Lower time complexity and tighter bound. In *Advances in Neural Information Processing Systems 25*, pages 647–655, 2012.
- [23] S. Wang and Z. Zhang. Improving cur matrix decomposition and the nystrom approximation via adaptive sampling. *Journal of Machine Learning Research*, 14:2729–2769, 2013.
- [24] S. Wang, Z. Zhang, and T. Zhang. Towards more efficient nystrom approximation and cur matrix decomposition. Technical report, 2015.
- [25] C.K. I. Williams and M. Seeger. Using the nystrom method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*.
- [26] T. Yang, L. Zhang, R. Jin, and S. Zhu. An explicit sampling dependent spectral error bound for column subset selection. In *International Conference on Machine Learning*, pages 135–143, 2015.
- [27] S. Yun, M. Ielarge, and A. Proutiere. Fast and memory optimal low-rank matrix approximation. In *Advances in Neural Information Processing Systems 28*, pages 3177–3185, 2015.
- [28] K. Zhang, I. Tsang, and J. Kwok. Improved nystrom low-rank approximation and error analysis. In *International conference on Machine learning*, pages 1232–1239, 2008.
- [29] T. Zhou and D. Tao. Godec: Randomized low-rank & sparse matrix decomposition in noisy case (2011). In *International Conference on Machine Learning*, 2011.

7 Proof of Theorem 1

In order to prove theorem 1, we need to partition the input matrix \mathbf{A} into small, equal-sized blocks. Note that this new partition will be determined based on the clusters defined on Section 3.1. Remind that the rows in \mathbf{P} are grouped into k clusters with cluster size $T_{(i)}^r$. We add virtual instances to \mathbf{P} such that all clusters have the same size, i.e., $T^r = \max T_{(i)}^r$. The virtual instances added to the q th cluster are chosen as the q th representative in \mathbf{Z}^r , therefore they induce no extra quantization errors. Then, we can re-partition \mathbf{P} into T^r partitions each containing exactly k instances. We use \mathcal{I}_i^r to denote the indexes of these partitions for $i = 1, 2, \dots, T^r$; similarly rows in \mathbf{Q} are also completed by virtual rows, falling in partitions \mathcal{I}_j^c for $j = 1, 2, \dots, T^c$. Equivalently, \mathbf{A} is augmented into kT^r -by- kT^c matrix, forming a number of $T^c T^r$ blocks each is of size k -by- k . The approximation error on each of these blocks will be quantized as follows.

Proposition 1 *Let the input matrix \mathbf{A} be augmented as described above, and the resultant decompositions \mathbf{P} and \mathbf{Q} are re-organized into T^r and T^c equal-sized groups, indexed by \mathcal{I}_i^r for $i = 1, 2, \dots, T^r$, and \mathcal{I}_j^c for $j = 1, 2, \dots, T^c$. Then the approximation on each block defined by indexes \mathcal{I}_i^r and \mathcal{I}_j^c is as follows.*

$$\frac{1}{\sqrt{k\theta}} \left\| \mathbf{A}_{[\mathcal{I}_i^r, \mathcal{I}_j^c]} - \mathbf{C}_{[\mathcal{I}_i^r, :]} \mathbf{W}^\dagger \mathbf{R}_{[\mathcal{I}_j^c, :]}^\top \right\|_F \leq \sqrt{e_j^r + e_j^c} + \sqrt{e_i^r} + \sqrt{e_i^c} + \sqrt{k\theta e_i^r e_j^c} \|\mathbf{W}^\dagger\|_F. \quad (7)$$

Here $e_i^r = \sum_{l \in \mathcal{I}_i^r} \|\mathbf{P}_{[l, :]} - \mathbf{Z}_{[s(l), :]}^r\|^2$ is the error of encoding rows of \mathbf{P} (specified by \mathcal{I}_i^r) with representative \mathbf{Z}^r via the mapping s^r . Similarly, $e_i^c = \sum_{l \in \mathcal{I}_i^c} \|\mathbf{P}_{[l, :]} - \mathbf{Z}_{[s(l), :]}^c\|^2$ is the encoding error of encoding rows in \mathbf{Q} (specified by \mathcal{I}_j^c) with representative \mathbf{Z}^c .

We first establish some basic equalities to use in the proof. Define $\langle \mathbf{X}, \mathbf{Y} \rangle = \mathbf{X}\mathbf{Y}^\top$ for matrices \mathbf{X}, \mathbf{Y} with proper dimensions.

$$\begin{aligned} \mathbf{C}_{[\mathcal{I}_i^r, :]} &= \langle \mathbf{P}_{[\mathcal{I}_i^r, :]}, \mathbf{Q}_{[\mathcal{Z}^c, :]} \rangle, \\ \mathbf{R}_{[\mathcal{I}_j^c, :]} &= \langle \mathbf{P}_{[\mathcal{I}_j^c, :]}, \mathbf{Q}_{[\mathcal{Z}^r, :]} \rangle, \\ \mathbf{A}_{[\mathcal{I}_i^r, \mathcal{I}_j^c]} &= \langle \mathbf{P}_{[\mathcal{I}_i^r, :]}, \mathbf{Q}_{[\mathcal{I}_j^c, :]} \rangle, \\ \mathbf{W} &= \langle \mathbf{P}_{[\mathcal{Z}^r, :]} , \mathbf{Q}_{[\mathcal{Z}^c, :]} \rangle. \end{aligned}$$

Here we have used the transposed version of \mathbf{R} for convenience of proof. In other words \mathbf{R} will be an $n \times k$ matrix, which is the transpose of its counterpart in theorem 1 or the CUR decomposition. The change of representation won't affect the correctness of our proofs.

We also define the following difference matrices

$$\begin{aligned} \Delta_C &= \mathbf{C}_{[\mathcal{I}_i^r, :]} - \mathbf{W}, \\ \Delta_R &= \mathbf{R}_{[\mathcal{I}_j^c, :]} - \mathbf{W}, \\ \Delta_A &= \mathbf{A}_{[\mathcal{I}_i^r, \mathcal{I}_j^c]} - \mathbf{W}, \end{aligned}$$

Proposition 2 *Let $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2$ be $1 \times d$ vectors (to be consistent with our definitions), and let $\phi(\cdot, \cdot)$ be the inner product between two such vectors, then we have the following inequality*

$$(\phi(\mathbf{x}_1, \mathbf{y}_1) - \phi(\mathbf{x}_2, \mathbf{y}_2))^2 \leq \theta \cdot (\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \|\mathbf{y}_1 - \mathbf{y}_2\|^2). \quad (8)$$

Proof 1 *Using the Lagrangian mean-value theorem, we can have*

$$\begin{aligned} (\phi(\mathbf{x}_1, \mathbf{y}_1) - \phi(\mathbf{x}_2, \mathbf{y}_2))^2 &= (\phi'(\xi)(\mathbf{x}_1 \mathbf{y}_1' - \mathbf{x}_2 \mathbf{y}_2'))^2 \\ &= \phi'(\xi)^2 (\mathbf{x}_1 \mathbf{y}_1' - \mathbf{x}_1 \mathbf{y}_2' + \mathbf{x}_1 \mathbf{y}_2' - \mathbf{x}_2 \mathbf{y}_2')^2 \\ &= \phi'(\xi)^2 (\mathbf{x}_1 (\mathbf{y}_1 - \mathbf{y}_2)' + (\mathbf{x}_1 - \mathbf{x}_2) \mathbf{y}_2')^2 \\ &\leq 2\phi'(\xi)^2 ((\mathbf{x}_1 (\mathbf{y}_1 - \mathbf{y}_2)')^2 + (\mathbf{x}_1 - \mathbf{x}_2) \mathbf{y}_2'^2) \\ &\leq \theta (\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \|\mathbf{y}_1 - \mathbf{y}_2\|^2) \end{aligned}$$

where $\theta = 2f'(\xi)^2$. This completes the proof of Proposition 2.

Using Proposition 2, we can bound the norms of the difference matrices as follows,

$$\begin{aligned}
\|\Delta_A\|_F^2 &= \|\mathbf{A}_{[\mathcal{I}_i^r, \mathcal{I}_j^c]} - \mathbf{W}\|_F^2 \\
&= \left\| \langle \mathbf{P}_{[\mathcal{I}_i^r, :]}, \mathbf{Q}_{[\mathcal{I}_j^c, :]} \rangle - \langle \mathbf{P}_{[\mathcal{Z}^r, :]}, \mathbf{Q}_{[\mathcal{Z}^c, :]} \rangle \right\|_F^2 \\
&= \sum_{p,q=1}^k \left[\phi \left(\mathbf{P}_{[\mathcal{I}_i^r(p), :]}, \mathbf{Q}_{[\mathcal{I}_j^c(q), :]} \right) - \phi \left(\mathbf{P}_{[\mathcal{Z}^r(p), :]}, \mathbf{Q}_{[\mathcal{Z}^c(q), :]} \right) \right]^2 \\
&\leq \theta \sum_{p,q=1}^k \left(\left\| \mathbf{P}_{[\mathcal{I}_i^r(p), :]} - \mathbf{P}_{[\mathcal{Z}^r(p), :]} \right\|^2 + \left\| \mathbf{Q}_{[\mathcal{I}_j^c(q), :]} - \mathbf{Q}_{[\mathcal{Z}^c(q), :]} \right\|^2 \right) \\
&= k\theta \left(\sum_{p=1}^k \left\| \mathbf{P}_{[\mathcal{I}_i^r(p), :]} - \mathbf{P}_{[\mathcal{Z}^r(p), :]} \right\|^2 + \sum_{q=1}^k \left\| \mathbf{Q}_{[\mathcal{I}_j^c(q), :]} - \mathbf{Q}_{[\mathcal{Z}^c(q), :]} \right\|^2 \right) \\
&= k\theta (e_i^r + e_j^c).
\end{aligned} \tag{9}$$

Here we have used the pre-defined relation $s^r(\mathcal{I}_i^r(p)) = p$, and $s^c(\mathcal{I}_j^c(q)) = q$, since the partition index \mathcal{I}_i^c and \mathcal{I}_j^c has the corresponding representative set \mathcal{Z} .

Similarly, we have

$$\begin{aligned}
\|\Delta_C\|_F^2 &= \|\mathbf{C}_{[\mathcal{I}_i^r, :]} - \mathbf{W}\|_F^2 \\
&= \left\| \langle \mathbf{P}_{[\mathcal{I}_i^r, :]}, \mathbf{Q}_{[\mathcal{Z}^c, :]} \rangle - \langle \mathbf{P}_{[\mathcal{Z}^r, :]}, \mathbf{Q}_{[\mathcal{Z}^c, :]} \rangle \right\|_F^2 \\
&\leq \theta \sum_{p,q=1}^k \left(\left\| \mathbf{P}_{[\mathcal{I}_i^r(p), :]} - \mathbf{P}_{[\mathcal{Z}^r(p), :]} \right\|^2 \right) \\
&= \theta k e_i^r.
\end{aligned} \tag{10}$$

and

$$\|\Delta_R\|_F^2 \leq \theta k e_j^c. \tag{11}$$

By using (9), (10), and (11), we have

$$\begin{aligned}
\left\| \mathbf{A}_{[\mathcal{I}_i^r, \mathcal{I}_j^c]} - \mathbf{C}_{[\mathcal{I}_i^r, :]} \mathbf{W}^\dagger \mathbf{R}_{[\mathcal{I}_j^c, :]}^\top \right\|_F &= \|(\mathbf{W} + \Delta_A) - (\mathbf{W} + \Delta_R) \mathbf{W}^\dagger (\mathbf{W} + \Delta_C)\| \\
&= \|\Delta_A - \Delta_R - \Delta_C - \Delta_R \mathbf{W}^\dagger \Delta_C\|_F \\
&\leq \|\Delta_A\|_F + \|\Delta_C\|_F + \|\Delta_A\|_F \|\Delta_C\|_F \|\mathbf{W}^\dagger\|_F \\
&= \sqrt{k\theta} \left(\sqrt{e_j^r + e_j^c} + \sqrt{e_i^r} + \sqrt{e_i^c} + \sqrt{k\theta e_i^r e_j^c} \|\mathbf{W}^\dagger\|_F \right).
\end{aligned}$$

This completes the proof of Proposition 1.

With this proposition, and by using the inequality $\sum_i^n \sqrt{x_i} \leq \sqrt{n \sum_i x_i}$, the overall approximation error can be bounded as follows

$$\begin{aligned}
\frac{1}{\sqrt{k\theta}} \|\mathbf{A} - \mathbf{C} \mathbf{W}^\dagger \mathbf{R}^\top\|_F &\leq \frac{1}{\sqrt{k\theta}} \sum_{i=1}^{T^r} \sum_{j=1}^{T^c} \left\| \mathbf{A}_{[\mathcal{I}_i^r, \mathcal{I}_j^c]} - \mathbf{C}_{[\mathcal{I}_i^r, :]} \mathbf{W}^\dagger \mathbf{R}_{[\mathcal{I}_j^c, :]}^\top \right\|_F \\
&\leq \sqrt{T^c} \sum_i \sqrt{\sum_j (e_j^c + e_j^r)} + \sqrt{T^r} \sum_j \sqrt{\sum_i e_i^r} + \sqrt{T^c} \sum_i \sqrt{\sum_j e_j^c} \\
&\quad + \sqrt{k\theta} \|\mathbf{W}^\dagger\|_F \sqrt{T^c} \sum_i \sqrt{\sum_j e_i^r e_j^c} \\
&\leq \sqrt{T^c T^r (T^r e^c + T^c e^r)} + T^c \sqrt{T^r e^r} + T^r \sqrt{T^c e^c} + \sqrt{k\theta} \sqrt{T^c T^r e^c e^r} \|\mathbf{W}^\dagger\|_F \\
&\leq \sqrt{3(T^c + T^r)(e^c + e^r)} + \sqrt{k\theta} \sqrt{T^c T^r e^c e^r} \|\mathbf{W}^\dagger\|_F.
\end{aligned}$$

By using $T = \max(T^r, T^c)$, we can easily prove Theorem 1.

8 Stability Issue

In this section, we will provide a detailed example showing that: (1) Nyström method is stable on PSD matrices by selecting the same subset of rows and columns in the approximation process; (2) both the pseudo-skeleton method and the sketch-CUR method are unstable on PSD matrices if they perform sampling of the rows and columns independently¹. This observation shows that the symmetric structure of PSD matrices allows one to impose useful constraints on the sampling, so as to guarantee the stability of low-rank approximation. However, in general rectangular matrices, no such constraints are available, therefore the low-rank approximation can be much more difficult.

In Figure 1, we plot the approximation error of the three methods, namely sketch-CUR, pseudo-skeleton, and Nyström method versus the number (ratio) of singular vectors used in computing the pseudo-inverse. The input matrix is chosen as an RBF kernel matrix, which is known to be symmetric, positive semi-definite (PSD). As can be seen, both the sketch-CUR method and the Pseudo-skeleton method are unstable, in that their approximation error is quite sensitive to the number of singular vectors adopted. In comparison, the Nyström method is stable: the more singular vectors adopted in computing the pseudo-inverse, the better the performance.

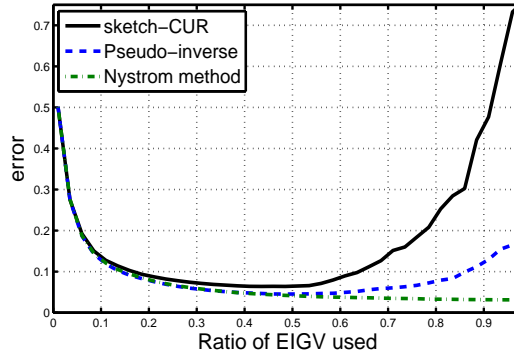


Figure 8: Approximation error versus the number of singular vectors used.

9 Orthogonalization Step of sketching Routine

The sketching routine in CABS (Algorithm 2) gives the following approximation

$$\mathbf{A} \approx \mathbf{U}\Sigma\mathbf{V}^\top. \quad (12)$$

Here $\mathbf{U} \in \mathbb{R}^{m \times r}$, $\Sigma \in \mathbb{R}^{k \times k}$ is a diagonal matrix, and $\mathbf{V} \in \mathbb{R}^{n \times k}$. In using this sketching results for subsequent steps, one has the option to further orthogonalize \mathbf{U} and \mathbf{V} , for example, when approximate leverage scores need to be computed. Empirically, we find that the orthogonalization step does not make a significant difference on the final performance, since \mathbf{U} and \mathbf{V} are already close to being orthogonal. The orthogonalization can be done as follows. Let the SVD decomposition of \mathbf{U} be $\mathbf{U} = \mathbf{U}_0 \Sigma_0 \mathbf{V}_0^\top$. Then perform another SVD decomposition $\Sigma_0 \mathbf{V}_0^\top \Sigma \mathbf{V}^\top = \mathbf{U}_1 \Sigma_1 \mathbf{V}_1^\top$. Finally, the orthogonalized left and right singular vector matrix would be $\mathbf{U}_0 \mathbf{U}_1$ and \mathbf{V}_1 , respectively, and the singular value matrix would be Σ_1 . In other words $\mathbf{A} \approx (\mathbf{U}_0 \mathbf{U}_1) \Sigma_1 (\mathbf{V}_1^\top)$. It's easy to verify that the computational cost is only $\mathcal{O}((m+n)k^2)$.

10 Proof of Theorem 2

First, note that the CABS algorithm uses only the k -dimensional embedding instead of the exact embedding as stated in theorem. The consequence is that the resultant error bound will be loosened by the trailing singular values of the

¹If they sample the same set of rows and columns, then they will be reduced to the Nyström method.

input matrix, as follows

$$\begin{aligned}
\|\mathbf{A} - \mathbf{C}\mathbf{W}^\dagger\mathbf{R}^\top\|_F &= \|\mathbf{A}_k - \mathbf{C}\mathbf{W}^\dagger\mathbf{R}^\top + \mathbf{A}_{\bar{k}}\|_F \\
&\leq \|\mathbf{A}_k - \mathbf{C}\mathbf{W}^\dagger\mathbf{R}^\top\|_F + \|\mathbf{A}_{\bar{k}}\|_F \\
&\leq \sqrt{T^c T^r} \left(\sqrt{3k\theta(e^r + e^c)(T^r + T^c)} + k\theta\sqrt{e^c e^r} \|\mathbf{W}^\dagger\|_F \right) + \|\mathbf{A}_{\bar{k}}\|_F \\
&= \mu\sqrt{e^r + e^c} + \nu\sqrt{e^r \cdot e^c} + \|\mathbf{A}_{\bar{k}}\|_F.
\end{aligned}$$

where

$$\begin{aligned}
\mu &= \sqrt{3k\theta T^c T^r (T^r + T^c)}, \\
\nu &= k\theta\sqrt{T^c T^r} \|\mathbf{W}^\dagger\|_F,
\end{aligned}$$

and $\|\mathbf{A}_{\bar{k}}\|_F = \sqrt{\sum_{i=k+1}^{\min(m,n)} \sigma_i^2}$ is a constant which is the l_2 -norm of the $\min(m, n) - k$ singular values. In case the singular-value spectrum decays rapidly, this constant can be quite small. In other words the error bound is only slightly loosened in case only (approximate) rank- k embeddings (instead of the exact embeddings) are used for the follow-up sampling.

In the following we will use the updated error bound for the pilot and follow-up sketching, as

$$\begin{aligned}
\Psi_p &= \mu_p \sqrt{e_p^r + e_p^c} + \nu_p \sqrt{e_p^r \cdot e_p^c} + \|\mathbf{A}_{\bar{k}}\|, \\
\Psi_f &= \mu_f \sqrt{e_f^r + e_f^c} + \nu_f \sqrt{e_f^r \cdot e_f^c} + \|\mathbf{A}_{\bar{k}}\|,
\end{aligned}$$

and

$$\begin{aligned}
\mu_p &= \sqrt{3k\theta T_p^c T_p^r (T_p^r + T_p^c)}, \quad \nu_p = k\theta\sqrt{T_p^c T_p^r} \|\mathbf{W}_p^\dagger\|_F, \\
\mu_f &= \sqrt{3k\theta T_f^c T_f^r (T_f^r + T_f^c)}, \quad \nu_f = k\theta\sqrt{T_f^c T_f^r} \|\mathbf{W}_f^\dagger\|_F.
\end{aligned}$$

Here the sub-index $\{p, f\}$ denotes the pilot and the follow-up step, and all parameters are defined in the same way as in Theorem 1. For example, $T_{p,f}^c$ and $T_{p,f}^r$ are the maximum cluster sizes in the column and row embeddings; $e_{p,f}^c$ and $e_{p,f}^r$ are the encoding errors for the column and row embeddings. The above relation holds because the random sampling in the pilot step can be deemed as equivalently running on the rank- k embeddings of the input matrix. This instantly gives the following guarantee

$$\begin{aligned}
\Delta_e^r &= e_p^r - e_f^r \geq 0, \\
\Delta_e^c &= e_p^c - e_f^c \geq 0.
\end{aligned}$$

Here Δ_e^r and Δ_e^c are exactly the drop of the encoding errors achieved by the k -means sampling algorithm in the follow-up sampling step. Next, we will show that, the drop of the error bounds from the pilot sketching step to the follow-up sketching step in CABS, can be exactly quantified by the drop of the encoding errors Δ_e^r and Δ_e^c .

We will also use the inequality $g(x) - g(y) \geq (x - y) \cdot g'(x)$ for the function $g(x) = \sqrt{x}$ and any pair of numbers $x \geq y \geq 0$. Namely $\sqrt{x} - \sqrt{y} \geq (x - y) \frac{1}{2\sqrt{x}}$. We make the realistic assumption that $T_p^r = T_f^r = T^r$, and $T_p^c = T_f^c = T^c$, since as we mentioned, the pilot random sampling and the follow-up k -means sampling can be deemed as running on the same, rank- k embedding of the input matrix. Namely the maximum cluster sizes in the two rounds of samplings will be the same. So we can safely write $\mu_p = \mu_f = \mu$. On other hand, we also assume that $\|\mathbf{W}_f^\dagger\|_F \leq \|\mathbf{W}_p^\dagger\|_F$. This is because the follow-up sampling using the k -means sampling will pick highly non-redundant rows and columns as the representatives, therefore the norm of the resultant intersection matrix $\|\mathbf{W}^\dagger\|_F$ will typically drop. In other words,

$$\begin{aligned}
\nu_p \sqrt{e_p^r \cdot e_p^c} - \nu_f \sqrt{e_f^r \cdot e_f^c} &= k\theta\sqrt{e_p^r \cdot e_p^c} \sqrt{T^c T^r} \|\mathbf{W}_p^\dagger\|_F - k\theta\sqrt{e_f^r \cdot e_f^c} \sqrt{T^c T^r} \|\mathbf{W}_f^\dagger\|_F \\
&\geq k\theta\sqrt{e_p^r \cdot e_p^c} \sqrt{T^c T^r} \|\mathbf{W}_p^\dagger\|_F - k\theta\sqrt{e_f^r \cdot e_f^c} \sqrt{T^c T^r} \|\mathbf{W}_p^\dagger\|_F \\
&= k\theta\sqrt{T^c T^r} \|\mathbf{W}_p^\dagger\|_F \left(\sqrt{e_p^r \cdot e_p^c} - \sqrt{e_f^r \cdot e_f^c} \right) \\
&\geq 0.
\end{aligned}$$

So we further bound the difference as follows

$$\begin{aligned}
\Psi_p - \Psi_f &= \left(\mu_p \sqrt{e_p^r + e_p^c} - \mu_p \sqrt{e_f^r + e_f^c} \right) + \left(\nu_p \sqrt{e_p^r \cdot e_p^c} - \nu_f \sqrt{e_f^r \cdot e_f^c} \right) \\
&\geq \mu \left(\sqrt{e_p^r + e_p^c} - \sqrt{e_f^r + e_f^c} \right) + k\theta \sqrt{T^c T^r} \|\mathbf{W}_p^\dagger\|_F \left(\sqrt{e_p^r \cdot e_p^c} - \sqrt{e_f^r \cdot e_f^c} \right) \\
&\geq \mu \frac{1}{2\sqrt{e_p^r + e_p^c}} ((e_p^r - e_f^r) + (e_p^c - e_f^c)) + k\theta \sqrt{T^c T^r} \|\mathbf{W}_p^\dagger\|_F \frac{1}{\sqrt{e_p^r \cdot e_p^c}} (e_p^r \cdot e_p^c - e_f^r \cdot e_f^c) \\
&= \sqrt{3k\theta T^c T^r (T^r + T^c)} \frac{1}{2\sqrt{e_p^r + e_p^c}} (\Delta_e^r + \Delta_e^c) \\
&\quad + k\theta \sqrt{T^c T^r} \|\mathbf{W}_p^\dagger\|_F \frac{1}{\sqrt{e_p^r \cdot e_p^c}} (\Delta_e^r e_f^c + \Delta_e^c e_f^r + \Delta_e^r \Delta_e^c).
\end{aligned}$$

This completes the proof of Theorem 2.